TOPICS IN GENERALIZED LEAST SQUARES SIGNAL ESTIMATION

P. Swerling

September 1965

P-3007-1

TOPICS IN GENERALIZED LEAST SQUARES SIGNAL ESTIMATION

P. Swerling
Consultant to The RAND Corporation, Santa Monica, California

Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

I. INTRODUCTION

Among approaches to the estimation of signals, or signal parameters, in the presence of noise, the following can be distinguished:

a. Approaches in which a-priori statistics are not considered to be associated with the parameters to be estimated, versus approaches in which a-priori statistics are associated with the unknown parameters.

Typical methods utilized in connection with the former approach are maximum likelihood,[1,2] minimum variance unbiased,[3,4] and use of Cramer-Rao, Barankin, or other lower bounds.[3,4]

In cases where a-priori statistics are associated with the unknown parameters, typical approaches involve a-posteriori probability[5] and decision theoretic[6] or Bayes optimum methods.*

*One of the reviewers has pointed out that the case where, in the above terminology, a-priori statistics are not associated with the parameters, can be interpreted in the sense that a-priori statistics are implicitly assumed having approximately uniform probability density over a very wide range; and in fact he considers such an interpretation to be a necessity in relating statistical criteria to the real world. On the other hand, the data smoothing methods to be treated herein have been widely used often without any such interpretation being made by the user.

For purposes of this paper, which largely concerns the various mathematically equivalent forms which can be taken by maximum likelihood, maximum a-posteriori, and least squares estimation procedures, the author believes this issue to be of secondary importance, and prefers to continue using a terminology according to which a-priori statistics may or may not be associated with given parameters.

The reader is free to interpret all statements in the following, to the effect that "no a-priori statistics" are associated with parameters, to be a short way of saying that a-priori statistics are assumed which have uniform density over a wide range.

b. Approaches in which the signals are regarded as deterministic except for some set of initially unknown parameters, typical of which are most treatments of the estimation of radar waveform parameters,[1,2] versus approaches in which the signals are regarded as random processes, as in Wiener filtering theory and its generalizations.[7,8]

One may also distinguish cases in which the number of initially unknown parameters is finite (as in most radar waveform estimation applications), and those in which there is an infinite set of initially unknown parameters (examples of which will be given in this paper).

The case in which the signals are regarded as random processes can be considered equivalent to that in which there is an infinite set of unknown parameters (since the signal process can be represented in terms of such a set) and in which a-priori statistics are associated with these parameters. In cases where there are only a finite number of unknown parameters, but a-priori statistics are associated with them, the signal can also be regarded as a random process, albeit one whose sample space is finite dimensional.

Section II of this paper is devoted to treating the maximum likelihood, maximum a-posteriori, and least squares approaches to these various cases on a unified basis. In Section II. 1, it is shown that the Maximum-A-Posteriori (MAP) estimate for the case where a-priori statistics are associated with the unknown parameters is under certain conditions equivalent to a Maximum Likelihood (ML) estimate in an equivalent problem in which the a-priori statistics associated with the parameters are regarded as providing additional equivalent observations of the parameters (this is not the same as the well known fact that the ML estimate is equal to the MAP estimate if the a-priori pdf's of the parameters are uniform, although the latter is a special case).

We also define mixed ML-MAP estimates for cases in which some but not all of the parameters are assumed to have a-priori statistical distributions; equivalent formulations are then given where these same estimates appear as pure ML estimates or alternatively as pure MAP estimates.

In Section II. 2, attention is turned to the main subject matter of this paper: the class of generalized least squares estimation procedures, which give the ML-MAP estimates if the statistics are Gaussian, but which remain good estimates even for non-Gaussian statistics. These are formulated for the case where a-priori statistics associated with the unknown parameters, or with some subset thereof, are regarded as providing the equivalent of additional observations, and a first order error analysis is given, from which the estimation error statistics are then derived. A result is also proved according to which parameters having a-priori statistics can be under certain conditions considered equivalent to additional noise, insofar as concerns estimation of other parameters.

The foregoing analysis is then applied in Section II. 3 to linear minimum mean square error filtering theory, which, as is shown, can be regarded as an application of parameter estimation with an infinite number of parameters, with a-priori statistics associated with them, the a-priori statistics in turn being regarded as providing the equivalent of additional observational data.

In Section II. 3 we also apply the results of Section II. 2 to prove that additive noise may itself be considered to contribute an additional infinity of parameters to be estimated jointly with the signal parameters (the noise statistics being considered to contribute an infinite number of additional equivalent "observations"). The signal parameter estimates are shown to be in certain cases the same, whether the noise is considered as noise or as an additional number of parameters to be jointly estimated with the signal parameters.

Section III is devoted to the application of the results of Section II to the problem of ob aining recursive solutions to certain very general forms of the signal estimation problem. These recursive solutions are of the type defined and analyzed by Swerling(9,10) and further investigated by Kalman,(11) Kalman and Bucy,(12) Blum,(13) and others.

Recursive solutions are derived in Section III. 2 for the case where the observed signal consists of the sum of K random processes called "signal" processes added to another random process called the "noise" process. The only assumption made regarding the signal processes is that they be continuous in the mean; the only assumption regarding the noise process is that it consist of the sum of two components, one of which is continuous in the mean, the other being white noise.

The problem is reduced to the pure white noise case by regarding the non-white noise component as an additional process to be estimated, and further regarding all the signal processes, as well as the non-white component of the noise process, as equivalent to an infinite number of parameters to be jointly estimated. The recursive techniques defined by Swerling(9) are then applied directly.

The result is a set of simultaneous, non-linear partial differential equations, the solution to which gives the desired recursive solution to the optimum linear filtering problem. However, this is not as bad as it sounds, since the recursive solution to the filtering problem results directly from, and is in fact identical with, the process of building up the solutions to the set of partial differential equations from the initial conditions.

Section III. 3 is devoted to extending these results to cases where some components of the noise may be non-additive.

Section IV contains some discussion of additional problems and applications suggested by the results of previous sections.

There is also an appendix which treats an example of an estimation problem involving an infinite number of unknown parameters; in one version, the problem can be solved without associating a-priori statistics with the unknown parameters, while a slightly modified version cannot be properly formulated or solved without associating a-priori statistics with the unknown parameters.

This paper is not completely self-contained, since heavy reliance is placed on the results of References 2 and 9. The most important formulas required are reproduced, but the discussion of a number of points is abbreviated, with reference being made to further discussion in the cited papers.

II. ML, MAP, and Generalized Least Squares Estimates

II. 1. Maximum Likelihood and Maximum A-Posteriori Estimates

Let S represent a set of observational data, and let x represent a parameter (possibly multiply, or even infinite, dimensional) upon which the probability distribution of S depends. Suppose that x has an a-priori probability density function p(x) associated with it.* We will suppose that the joint probability density function of S and x exists and is denoted p(S,x). The existence of the conditional probability densities $p(S|x)$ and $p(x|S)$ is also assumed. Then, the maximum a-posteriori (MAP) estimate $\hat{x}$ of x is that value of x which maximizes $p(x|S)$. (In the following, the symbol S will be used to denote both the observed value of a random variable and the argument of various pdf's; also, the sympol p will be used to denote a variety of pdf's with, hopefully, confusion avoided by the fact that the argument of p will indicate which pdf we are talking about.)

Now suppose that the a-priori pdf of x, p(x), is symmetrical about some point $\bar{x}$:

$$p(x) = f(x - \bar{x}) = f(-x + \bar{x}) \tag{1}$$

(If p(x) is also unimodal, then $\bar{x}$ would be the MAP estimate of x based on just the a-priori information.)

The following equivalent estimation problem can then be formulated:

$\bar{x}$ will be regarded as the observed value of an additional "equivalent" or "virtual" observation $S^{(e)}$. This additional

*Throughout the paper, pdf's are defined with respect to Lebesgue measure in finite-dimensional sample spaces. Many results are derived for infinite-dimensional sample spaces, but these are always derived by valid limiting processes from finite-dimensional approximations.

observation will be represented as

$$S^{(e)} = x + \delta\, S^{(e)} \qquad (2)$$

where $\delta\, S^{(e)}$ = equivalent observation error.

As stated, the observed value of the random variable in any case is assumed to be $\bar{x}$, so the value of the equivalent observation error is $\bar{x} - x$. Here, x may be a multi-component vector $(x_1, x_2, \ldots)$.

The random variable $S^{(e)}$ is assumed to have pdf characterized by the conditions:

$$E^{(e)} \left[S^{(e)}\right] = x \qquad (3)$$

$$\text{pdf of } \delta\, S^{(e)} = \text{a-priori p.d.f. of } \bar{x} - x$$

This amounts to saying

$$p^{(e)} \left[S^{(e)} \mid x\right] = f\left[S^{(e)} - x\right] = f\left[x - S^{(e)}\right] \qquad (4)$$

In this equivalent problem, we regard x as a constant, and $S^{(c)}$ as a random variable with pdf given by (4). It is also assumed that the equivalent random variable $\delta\, S^{(e)}$ is statistically independent of S.

Now, the maximum likelihood estimate of x for this equivalent problem is obtained as tollows. The likelihood function is

$$p\left[S,\ S^{(e)}\ \middle|\ x\right] = p\left[S\ \middle|\ x\right]\ p^{(e)}\left[S^{(e)}\ \middle|\ x\right] \tag{5}$$

$$= \frac{p(x\ |\ S)\ p^{(e)}\left[S^{(e)}\ \middle|\ x\right]\ q(S)}{p(x)}$$

where $\qquad q(S) = \int p(S,\ x)\ d\ x$

The maximum likelihood estimate of x is obtained by substituting the observed values of S and $S^{(e)}$ into (5) and then maximizing with respect to x. However, the observed value of $S^{(e)}$ is $\bar{x}$. Thus, when $\bar{x}$ is substituted for $S^{(e)}$ in (5), and use is then made of (4) and (1), we find that the maximum likelihood estimate of x is obtained by maximizing $p(x\ |\ S)$ and is thus equal to the MAP estimate for the original problem.

We can also define mixed ML-MAP estimates for the case where some but not all of the unknown parameters have a-priori statistics. Suppose we write

$$x = (u,\ v) \tag{6}$$

(each of u and v may be multi-dimensional).

Also suppose that an a-priori pdf $p(u\ |\ v)$ is associated with u (possibly the a-priori pdf of u depends on v as a parameter) but no a-priori statistics are associated with v. Then, the mixed ML-MAP estimate of x is defined to be

$$\hat{x} = \text{ML-MAP estimate of } x \qquad (7)$$

$$= \text{value which maximized } p(S \mid x)\, p(u \mid v)$$

We can define $\hat{x}$ also as a pure MAP estimate by associating with v an a-priori pdf p(v) which is uniform over an extremely wide range of values. If the resulting pdf for (u, v) = x is then denoted p(u, v) = p(x), then $\hat{x}$ is the value which maximizes p(S | x) p(x).

$\hat{x}$ can also be represented as a pure ML estimate by considering the a-priori statistics associated with u to be equivalent to providing an additional "virtual" observation $S^{(e)}$ whose observed value is $\bar{u}$ and whose pdf is given by

$$E^{(e)}\left[S^{(e)} \mid x\right] = u \qquad (8)$$

$$p^{(e)}\left[S^{(e)} \mid x\right] = f\left[S^{(e)} - u \mid v\right] \qquad (9)$$

$$= f\left[u - S^{(e)} \mid v\right]$$

where it has been assumed that p(u | v) is of the form

$$p(u \mid v) = f\left[u - \bar{u} \mid v\right] = f\left[\bar{u} - u \mid v\right] \qquad (10)$$

$\hat{x} = (\hat{u}, \hat{v})$ is then obtained by maximizing, with respect to (u, v), the quantity

$$p^{(e)} \left[S, S^{(e)} \mid u, v\right] = p(S \mid u, v)\, p^{(e)} \left[S^{(e)} \mid u, v\right] \qquad (11)$$

with the value $\bar{u}$ substituted for $S^{(e)}$ and the actual observational data substituted for S.

In the rest of this paper, when we adjoin equivalent observational data, in the above-described manner, to the actual observed data, in order to obtain an ML estimate equal to the original MAP or mixed ML-MAP estimate, we will call these "virtual" or "equivalent" observations as distinguished from the "actual" observations. Such "equivalent observations" may be regarded as the parameter estimates which would be made if there were no actual observed data but only a-priori statistics for the parameters; they will generally be denoted by a superscript "e". The author has found the concept of virtual observations a convenient way of looking at things in many applications, but that is all that is claimed for it.

II. 2. Generalized Least Squares Estimates

Attention will now be turned to a class of generalized least squares estimates analyzed extensively by Swerling.(2,9)

Suppose the observational data is given by

$$S_\mu = f_\mu(x_1, x_2, \ldots, t_\mu) + \mathcal{E}_\mu \qquad (12)$$

where $S_\mu = \mu^{th}$ observation (a real scalar)

$f_\mu(x, t_\mu) = f_\mu(x_1, x_2, \ldots, t_\mu)$ = value of μ^{th} observation in the absence of observation error

ε_μ = observation error for the μ^{th} observation

$x = (x_1, x_2, \ldots)$ = finite or infinite set of unknown parameters (real scalars), not yet assumed to have a-priori statistics.

t_μ = time of μ^{th} observation (assumed known)

Consider the class of estimation procedures defined as follows: $\hat{x}_i$ are estimates obtained by minimizing, with respect to $(x_1, x_2, \ldots)$ the quantity

$$Q = \sum_{\mu,\nu} \eta_{\mu\nu} \left[S_\mu - f_\mu(x, t_\mu)\right]\left[S_\nu - f_\nu(x, t_\nu)\right] \tag{13}$$

where $(\eta_{\mu\nu})$ = arbitrary symmetric positive definite matrix (not necessarily the inverse covariance matrix of $\{\varepsilon_\mu\}$; see below)

Before proceeding, several comments should be made:

a) The subscript μ on f_μ indicates that the functional dependence of the observations on the unknown parameters may differ for each observation. However, f_μ is assumed to be a known function. (If f_μ is not exactly known, this can still be represented by assuming f_μ to be known and compensating for this assumption by adding an equivalent observation error.[9])

b) The functions f_μ need not depend explicitly on t_μ. Also, t_μ may be replaced by one or more spatial or space-time parameters, or for that matter, by an element of an abstract parameter set. However, for each observation, t_μ (or the abstract parameter replacing it) is assumed known. If in a practical case t_μ is not perfectly known, we can still assume it to be known and compensate for this assumption by adding an equivalent observation error.

c) In most cases of interest, one or more components of the observational data may consist of functions of a continuous time (or other) parameter, of the form

$$S(t) = g(t, x_1, x_2, \ldots) + \mathcal{E}(t), \quad t \in (\tau_1, \tau_2) \tag{14}$$

In such a case, the term in Q contributed by such observations must be understood as the limiting form of finite sums. Such limiting forms can be well-defined and are often expressible as integrals.(1,2,4,14) (More precisely, this applies to those terms of Q remaining after terms which are independent of x are subtracted.)

For example, we could use discrete sampled times $\{t_\mu\}$ and have

$$\begin{aligned} S_\mu &= S(t_\mu) \\ f_\mu(x,t_\mu) &= g(t_\mu,x) \\ \mathcal{E}_\mu &= \mathcal{E}(t_\mu) \\ \eta_{\mu\nu} &= \left[\Phi(t_\mu,\ t_\nu)\right]^{-1} \end{aligned} \tag{15}$$

where Φ is a covariance function (not necessarily that of $\mathcal{E}(t)$; see below).

The corresponding term of Q is the limit as the set $\{t_\mu\}$ becomes dense in (τ_1, τ_2).

Alternatively one can consider the observed data to be the coefficients of S(t) with respect to a complete orthonormal set of functions $\{\Psi_\mu(t)\}$ over (τ_1, τ_2):

$$S_\mu = \int_{\tau_1}^{\tau_2} S(t)\, \Psi_\mu(t)\, dt \tag{16}$$

$$f_\mu(x) = \int_{\tau_1}^{\tau_2} g(t, x)\, \Psi_\mu(t)\, dt \tag{17}$$

(This is an example where f_μ does not depend explicitly on t_μ.)

If $\varepsilon(t)$ is considered a sample function of a random process with covariance function Φ, then a convenient choice for $\{\Psi_\mu\}$ is the set of orthonormal eigenfunctions associated with Φ over (τ_1, τ_2).

With this understanding, we will continue to write Q as a double sum.

Now, this method of estimation results in maximum likelihood estimates under the following two conditions:

A. $\{\varepsilon_\mu\}$ are jointly Gaussian random variables.

B. $\{\varepsilon_\mu\}$ have zero means, and $(\eta_{\mu\nu})$ is the inverse of the covariance matrix of the variables $\{\varepsilon_\mu\}$.

(For continuous time data, the appropriate limiting statements are understood.)

If these conditions hold, certain statements can be made about the optimality of the resulting estimates in the sense that they are minimum variance unbiased estimates, either precisely or asymptoticallv (see below). However, it is worth emphasizing that the estimates may still be good ones even if one or both of these conditions fail.

For example, if the statistics are not Gaussian, but condition B is satisfied, the variances of $(\hat{x}_i - x_i)$ are in many cases not affected; the only thing affected is what type of statements can be made about the optimality of the estimates.[2]

If condition B is not satisfied, there will in general be some degradation in estimation accuracy. However, it may be desirable from other points of view to use smoothing matrices ($\eta_{\mu\nu}$) not satisfying condition B. For example, it may be convenient for computational purposes to use a diagonal smoothing matrix even if the observation errors are correlated; or, one may not know the error covariances exactly. The increased computational convenience may be worth the decrease in accuracy. In any event, the degradation in accuracy caused by failure of condition B can usually be computed, as is described in the section of Reference 2 entitled "Mismatched Processing of Received Signals." Reference 2 also treats the degradation in estimation accuracy caused if the functions f_μ utilized in Q do not exactly describe the true dependence of the observations (in the absence of observation error) on the parameters x_1, x_2, ...

Thus, we will adopt the point of view that the method of minimizing Q defines estimates which may be good estimates whether or not conditions A and B are satisfied; in the special case where they are satisfied, the estimates are ML.

If the errors ε_μ are sufficiently small, the estimates $\hat{x}_i$ resulting from minimization can be written to first order as (2,9):

$$\hat{x}_i - x_i = \sum_\mu \gamma_{i\mu}(x)\, \varepsilon_\mu + \text{higher order terms} \tag{18}$$

$$\gamma_{i\mu}(x) = \sum_j \sum_\nu \left[B(x)\right]^{-1}_{ij} \eta_{\mu\nu} \frac{\partial f_\nu(x, t_\nu)}{\partial x_j} \tag{19}$$

$$B_{ij}(x) = \sum_{\mu,\nu} \eta_{\mu\nu} \frac{\partial f_\mu(x, t_\mu)}{\partial x_i} \frac{\partial f_\nu(x, t_\nu)}{\partial x_j} \tag{20}$$

It should be emphasized that Eqs. (18) - (20) depend only on the definition of the method of obtaining $\hat{x}_i$ by minimizing Q; they do not depend on any statistical interpretation of the quantities ε_μ nor on whether conditions A and B hold. (In Eqs. (18) - (20) and other places below, as will be clear from the context , $x = (x_1, x_2, \ldots)$ denote the true values of the parameters.)

If condition B is satisfied, and if the higher order terms can be neglected, then

$$E\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j)\right] = \left[B(x)\right]^{-1}_{ij} \tag{21}$$

We will now proceed to define a generalized least squares smoothing technique for the case where a-priori statistics are associated with some subset of the parameters x_1, x_2, ...

Let us suppose that I is a subset of (1,2,...), not necessarily a proper subset. Suppose that a joint a-priori statistical distribution is associated with those x_i for $i \in I$.

We will assume that Eq. (12) defines the actual observations. The generalized least squares estimates are obtained by adjoining to Q a term corresponding to the "equivalent observations" provided by the a-priori statistics.

In this case, Q is defined by

$$Q = \sum_{\mu,\nu} \eta_{\mu\nu}\left[S_\mu - f_\mu(x, t_\mu)\right]\left[S_\nu - f_\nu(x, t_\nu)\right] \tag{22}$$

$$+ \sum_{k,\ell\in I} \xi_{k\ell}\left[\bar{x}_k - x_k\right]\left[\bar{x}_\ell - x_\ell\right]$$

where

$(\xi_{k\ell})$ is an arbitrary symmetric positive definite matrix

$\bar{x}_k$, $k \in I$ are arbitrary constants

Thus, each component x_i of x which is associated with a-priori statistics is represented by an "observation" whose effective dependence on x is given by

$$f_i(x) = x_i, \ i \in I \tag{23}$$

and whose "observed value" is $\bar{x}_i$.

Of course, it is clear that in order to make good use of the a-priori information, $\xi_{k\ell}$ and $\bar{x}_k$ are not going to be completely arbitrary; in fact, $\bar{x}_k$ are going to be something like the means of the a-priori p d f's of x_k, and $\xi_{k\ell}$ are going to be related to the a-priori covariance matrix of $\{x_k - \bar{x}_k\}$. However, for present purposes we can regard any deviation of $(\xi_{k\ell})$ from the inverse covariance matrix of $\{x_k - \bar{x}_k\}$ as analogous to a deviation of $(\eta_{\mu\nu})$ from the inverse covariance matrix of $\{\varepsilon_\mu\}$ in the case of the actual observations; and any deviation of $\bar{x}_k$ from the means of the a-priori distributions as analogous to a deviation of the functions f_μ from those which truly describe t.ie dependence of the actual observations on x.

We wish to apply Eqs. (18) - (20) to obtain the first order dependence of $\hat{x}_i - x_i$ on $\{\varepsilon_\mu\}$ and $\{\bar{x}_k - x_k\}$, $k \in I$. To facilitate this, we introduce the following notation:

$$x = (u, v)$$
$$u = \{x_i\}, \ i \in I$$
$$\bar{u} = \{\bar{x}_i\}, \ i \in I$$
$$v = \{x_i\}, \ i \notin I$$

Then, to first order in $\{\varepsilon_\mu\}$ and $\{\bar{x}_k - x_k\}$, k ε I:

$$\hat{x}_i - x_i = \sum_\mu \gamma_{i\mu}(\bar{u}, v)\, \varepsilon_\mu + \sum_{k \epsilon I} \beta_{ik}(\bar{u}, v)\left[\bar{x}_k - x_k\right] \tag{24}$$

$$\gamma_{i\mu}(x) = \sum_j \sum_\upsilon \left[B(x)\right]^{-1}_{ij} \eta_{\mu\upsilon} \frac{\partial f_\upsilon(x, t_\upsilon)}{\partial x_j} \tag{25}$$

$$\beta_{ik}\ (x) = \sum_{j \epsilon I} \left[B(x)\right]^{-1}_{ij} \xi_{jk},\ k \epsilon I \tag{26}$$

$$B_{ij}(x) = \sum_{\mu,\upsilon} \eta_{\mu\upsilon} \frac{\partial f_\mu(x, t_\mu)}{\partial x_i} \frac{\partial f_\upsilon(x, t_\upsilon)}{\partial x_j} + \xi^*_{ij} \tag{27}$$

$$\xi^*_{ij} = \xi_{ij} \text{ if } i, j \epsilon I \tag{28}$$

$$= 0 \quad \text{if } i \notin I \text{ or } j \notin I$$

Now, consider the following conditions:

A′: $\{\varepsilon_\mu\}$ are jointly Gaussian; and $\{x_k\}$, k ε I have an a-priori joint Gaussian p d f.

B′: $\{\varepsilon_\mu\}$ have means zero; $\{x_k\}$, k ε I have means $\bar{x}_k$; and $\{x_k - \bar{x}_k\}$, k ε I are a-priori uncorrelated with $\{\varepsilon_\mu\}$. Also, $\{\varepsilon_\mu\}$ have inverse covariance matrix $(\eta_{\mu\upsilon})$, and $\{x_k - \bar{x}_k\}$, k ε I have inverse covariance matrix $(\xi_{k\ell})$. For the sake of simplicity, $(\xi_{k\ell})$ is also assumed to be independent of $\{x_i\}$, $i \notin I$.

If A′ and B′ are satisfied, $\hat{x}_i$ are the ML-MAP estimates (for all i).

We will now use Eqs. (24) - (28) to determine the covariance matrix $E\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j)\right]$ of the estimation errors supposing that just B' holds. It is also assumed tha* the first order equation (24) is valid (i.e., that higher order terms are negligible).

The covariance matrix of the estimation errors $\hat{x}_i - x_i$ will be computed with respect to the statistical ensemble defined by the statistics of the actual observation errors $\{\varepsilon_\mu\}$ and the a-priori statistics of $\{u_i\}$, $i \in I$. This can be done simply by forming the products $(\hat{x}_i - x_i)(\hat{x}_j - x_j)$ from Eqs. (24) - (28), taking expected values with respect to the joint p.d.f. of $\{\varepsilon_\mu\}$, and then taking the expected value with respect to the a-priori joint p d f of $\{u_i\}$, $i \in I$. The result is the following:

$$E\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j)\right] = \left[B(\bar{u}, v)\right]^{-1}_{ij} \tag{29}$$

An important special case is that where B is independent of x, in which case the right side of Eq. (29) becomes simply B^{-1}_{ij} . Many important applications fall into this category.

The above equations also clarify the sense in which the adjoining of "equivalent" observations is actually equivalent to the a-priori statistics.

Let the expected value of $(\hat{x}_i - x_i)(\hat{x}_j - x_j)$, given that the true value is x, with respect to the (fictitious) statistical ensemble defined by the actual observations and equivalent observations, with x regarded as a constant, be denoted by $E^{(e)}\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j) \mid x\right]$.

Then, from Eq. (21), to first order,

$$E^{(e)}\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j) \mid x\right] = \left[B(x)\right]^{-1}_{ij} \tag{30}$$

Thus, Eq. (29) says that, *provided the first order expansion (24) holds*, for $(\bar{u}_k - u_k)$ as well as for $\{\mathcal{E}_\mu\}$

$$E\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j)\right] = E^{(e)}\left[\hat{x}_i - x_i)(\hat{x}_j - x_j) \mid \bar{u}, v\right] \tag{31}$$

It is also true that, to first order,

$$E^{(e)}\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j) \mid \bar{u}, v\right] \tag{32}$$

$$= \int E^{(e)}\left[(\hat{x}_i - x_i)(\hat{x}_j - x_j) \mid u, v\right] p(u)\, du$$

Thus, provided the first order expansions are valid, the "equivalence" of the a-priori statistics to the adjoining of "equivalent" observations applies not only in the sense that the MAP estimates in the original problem are equal to the ML estimates in the equivalent problem (which holds regardless of the validity of the first order expansions), but also in that the estimation error covariance matrix for the original problem can be derived from that for the equivalent problem via Eq. (31) or Eq. (32).

In the case where the matrix B is independent of x, then so also are both sides of Eq. (31).*

*The expectations in Eqs. (29), (31), and (32) are conditional on v having a definite value. If one wishes to interpret "no a-priori statistics" as implicitly meaning "uniform-density a-priori statistics," then it must be understood that these expectations are not taken over these uniform a-priori statistics of v (unless we are dealing with a case where B is independent of v).

Another useful form of the first order equations for $\hat{x}_i$ can be stated in the case where all components of x have a-priori statistics, i.c., where x = u.

From Eqs. (20), (21), (23) and (25) of Ref. 9 one can state that, to first order,

$$\hat{x}_i - \bar{x}_i = \sum_j \left[B(\bar{x})\right]^{-1}_{ij} \rho^*_j (\bar{x}) \tag{33}$$

$$\rho^*_j (\bar{x}) = \sum_{\mu, \upsilon} \eta_{\mu\upsilon} \frac{\partial f_\mu(\bar{x}, t_\mu)}{\partial x_j} r_\upsilon(\bar{x}, t_\upsilon) \tag{34}$$

$$r_\upsilon(\bar{x}, t_\upsilon) = S_\upsilon - f_\upsilon(\bar{x}, t_\upsilon) \tag{35}$$

Attention will now be turned to the statement of a result according to which, in some cases, parameters having a-priori statistics can be considered equivalent to additive noise, insofar as concerns estimation of other parameters; or conversely, roise can be considered to be represented by additional parameters to be estimated.

Our initial statement of this result can, however, be stated in a form which does not involve statistical concepts:

Let

$$x = (u, v) \tag{36}$$

(Here, the notation (u,v) does not have the same significance as before; a-priori statistics will be associated with some components of u and with all components of v.)

$$S_\mu = f_\mu(u,\ v,\ t_\mu) + \mathcal{E}_\mu = g_\mu(u,\ t_\mu) + h_\mu(v,\ t_\mu) + \mathcal{E}(t_\mu) \tag{37}$$

$$\begin{aligned} Q(u,\ v) = \sum_{\mu,\nu} & \left[S_\mu - g_\mu(u,\ t_\mu) - h_\mu(v,\ t_\mu)\right] \\ & \times \left[S_\nu - g_\nu(u,\ t_\nu) - h_\nu(v,\ t_\nu)\right] \eta_{\mu\nu} \\ & + \sum_{k,\ell} \lambda_{k\ell}(\bar{u}_k - u_k)(\bar{u}_\ell - u_\ell) + \sum_{k,\ell} \xi_{k\ell}(\bar{v}_k - v_k)(\bar{v}_\ell - v_\ell) \end{aligned} \tag{38}$$

In Eq. (38), the sum involving terms $\lambda_{k\ell}(\bar{u}_k - u_k)(\bar{u}_\ell - u_\ell)$ is extended over a subset (not necessarily proper) of the indices of $\{u_k\}$; i.e., $\lambda_{k\ell} = 0$ unless both k and ℓ belong to some subset I of the indices of u. However, $(\lambda_{k\ell})$ is assumed to be positive definite when k, ℓ are restricted to I. On the other hand, the matrix $(\xi_{k\ell})$ is assumed to be positive definite with k, ℓ ranging over all the indices of $\{v_k\}$.

Suppose that the estimates $\hat{u}$, $\hat{v}$ are obtained by finding those values which minimize Q(u, v) with respect to u and v.

Now consider

$$\begin{aligned} Q^*(u) = \sum_{\mu,\nu} & \eta^*_{\mu\nu} \left[S_\mu - g_\mu(u,\ t_\mu) - h_\mu(\bar{v},\ t_\mu)\right] \\ & \times \left[S_\nu - g_\nu(u,\ t_\nu) - h_\nu(\bar{v},\ t_\nu)\right] \\ & + \sum_{k,\ell} \lambda_{k\ell}(\bar{u}_k - u_k)(\bar{u}_\ell - u_\ell) \end{aligned} \tag{39}$$

where

$$\eta^* = (\eta^{-1} + \phi^*)^{-1} \tag{40}$$

and

$$\phi^*_{\mu\nu} = \sum_{k,\ell} \xi^{-1}_{k\ell} \frac{\partial h_\mu(\bar{v}, t_\mu)}{\partial v_k} \frac{\partial h_\nu(\bar{v}, t_\nu)}{\partial v_\ell} \tag{41}$$

(In Eq. (40), η^*, η, and Φ^* are matrices.)

Let $\hat{u}^*$ be the estimate of u obtained by minimizing $Q^*(u)$ with respect to u. Then,

$$\hat{u}^* = \hat{u}, \text{ to first order} \tag{42}$$

It is to be noted that the result stated in Eqs. (36) - (42) has been stated, and can be proved, entirely without recourse to statistical concepts or interpretations. Before outlining the proof, however, the statistical motivation will be described:

Suppose we now regard $\{\varepsilon_\mu\}$ as a random process with means zero and inverse covariance matrix $(\eta_{\mu\nu})$ and suppose

(a) a-priori statistics are associated with some of the components of u, having means $\bar{u}_k$ and inverse covariance matrix $(\lambda_{k\ell})$

(b) a-priori statistics are associated with all of the components of v, having means $\bar{v}_k$ and inverse covariance matrix $(\xi_{k\ell})$

(c) $\{\varepsilon_\mu\}$, $\{u_k\}$, and $\{v_k\}$ are statistically uncorrelated

Write the actual observations as

$$S_\mu = f_\mu(u, v, t_\mu) + \mathcal{E}_\mu = g_\mu(u, t_\mu) + h_\mu(v, t_\mu) + \mathcal{E}_\mu \tag{43}$$

$$\approx g_\mu(u, t_\mu) + h_\mu(\bar{v}, t_\mu) + \mathcal{E}_\mu + \sum_k (v_k - \bar{v}_k) \frac{\partial h_\mu(\bar{v}, t_\mu)}{\partial v_k}$$

If we regard $\mathcal{E}_\mu + \sum_k (v_k - \bar{v}_k) \frac{\partial h_\mu(\bar{v}, t_\mu)}{\partial v_k}$

as the "noise", then this noise will be a random process with mean zero and inverse covariance matrix η^*. Consequently, Eq. (42) can be interpreted as follows:

Suppose the original problem, in which u and v are to be jointly estimated, is replaced by another problem in which u only is to be estimated; v is eliminated from the problem, and also the virtual observations equivalent to the a-priori statistics of v are eliminated. The virtual observations associated with u are retained unaltered.

The actual observations are replaced 'y

$$S_\mu = f^*_\mu(u, t_\mu) + \mathcal{E}^*_\mu \tag{44}$$

where

$$f^*_\mu(u, t_\mu) = g_\mu(u, t_\mu) + h_\mu(\bar{v}, t_\mu) \tag{45}$$

and $\{\mathcal{E}^*_\mu\}$ is a random process with zero means and inverse covariance matrix η^*.

$\hat{u}^*$ is the estimate for u obtained in this second problem (i.e., the ML-MAP estimate if conditions A' and B' hold).

Equation (42) states that, to first order, if A' and B' hold, the ML-MAP estimate $\hat{u}$ in the first problem is equal to the ML-MAP estimate $\hat{u}^*$ in the second problem.

The proof involves some rather tedious matrix algebra. An outline is as follows:

a) If the functions g and h are reasonably well-behaved, the estimate $\hat{u}$ can be obtained as follows (assuming we can effectively restrict the problem to the immediate neighborhood of $\hat{u}$, $\hat{v}$):

First find, for any fixed u, the value $\hat{v}(u)$ which minimizes Q(u, v) with respect to v.

Then let

$$Q'(u) = Q\left[u, \hat{v}(u)\right] \tag{46}$$

Then, $\hat{u}$ is that value which minimizes $Q'(u)$.

The proof of Eq. (42) then consists in verifying that, to first order, $Q'(u) = Q^*(u)$.

II.3. The Linear Case with an Infinite Set of Parameters

The primary aim of this subsection is to provide an example of the foregoing analysis by applying the results of Section II. 2 to the linear case with an infinity of unknown parameters; as will be seen, this amounts to another way of looking at standard linear minimum mean square error filtering theory. First, brief discussions will be given of linearity vs. non-linearity, and of finite vs. infinite parameter sets.

The Linear Case

We will define the linear case to be that case in which the functions f_μ depend linearly on the unknown parameters. In general, this may be written

$$f_\mu(x, t_\mu) = \sum_i x_i \, g_{\mu i} \, (t_\mu) + h(t_\mu) \tag{47}$$

where $g_{\mu i}(t)$ and $h(t)$ are known functions.

It should be noted that the virtual observations which are equivalent to a-priori statistics are of the form given by Eq. (1) and are therefore automatically in the linear form. Thus, if the functions f_μ describing the actual observations satisfy Eq. (47), then all the observations, both actual and virtual, are of the linear form.

In such a case, the following statements can be made:

(a) The estimates $\{\hat{x}_i\}$, if conditions A′ and B′ are satisfied, are exact minimum variance unbiased estimates. If only B′ is satisfied, they are still exact minimum variance unbiased linear estimates.

In the case of non-linear dependence of f_μ on the parameters, one can still in many cases say that the estimates $\{\hat{x}_i\}$ are asymptotically minimum variance; this is discussed further shortly.

(b) All the results obtained "to first order" in Section II. 2 can now be said to hold exactly without restriction on the magnitudes of $\{\varepsilon_\mu\}$ and of $\{\bar{x}_i - x_i\}$. Also, the matrices B and partial derivatives $\partial f_\mu \mid \partial x_i$ are independent of the values of $\{x_i\}$.

In the non-linear case, for sufficiently small values of $\{\varepsilon_\mu\}$ and of $\{\bar{x}_i - x_i\}$, the problem can be linearized and the results obtained to first order will be approximately correct.

Actually, in many cases, this linearization will lead to correct results even in cases where the individual values of $\{\varepsilon_\mu\}$ and $\{\bar{x}_i - x_i\}$ are fairly large, provided the total (integrated) signal-to-noise ratio, in some appropriately defined sense, is large.(2) However, there are some subtle pitfalls connected with determining the requirements on output signal-to-noise ratio in order to ensure that the results obtained from the linearized problem are correct. This is discussed at some length in Ref. 2.

One can give the following heuristic condition for the signal-to-noise ratio required in order that the solutions obtained from the linearized problem be approximately correct.

Suppose that the true parameter values are denoted by $\{x_i\}$, and that there exists a region R containing $x = \{x_i\}$ such that, for all x', x'' in R, and all μ,

$$f_\mu(x'', t_\mu) = f_\mu(x', t_\mu) + \sum_i (x_i'' - x_i') \frac{\partial f_\mu(x', t_\mu)}{\partial x_i} \tag{48}$$

$$+ \text{ remainder}$$

where the remainder term is negligible within R.

Also suppose that the output signal-to-noise ratio is sufficiently high so that, with probability approaching unity, a preliminary estimate $\hat{x}_o$ can be obtained with $\hat{x}_o \in R$.

Then, with probability equal essentially to unity, one can replace the original problem with the linearized problem in which the observations are replaced by $S_\mu - f_\mu(\hat{x}_o\ ,\ t_\mu)$; the parameters to be estimated are replaced by $\{x_i - \hat{x}_{o\,i}\}$; and the functions f_μ are replaced by

$$f_\mu^{\,*} = \sum_i \left[x_i - \hat{x}_{o\,i}\right] \frac{\partial\, f_\mu(\hat{x}_o\ ,\ t_\mu)}{\partial\, x_i} \tag{49}$$

Thus, the condition is that the signal-to-noise ratio be sufficiently high that the problem can, with probability essentially equal to unity, be confined to a region R around x where the variation of f_μ with $\{x_i\}$ is linear except for a negligible remainder.

Finite vs. Infinite Parameter Sets

In typical cases, the estimation errors due to observation errors in least-squares smoothing methods of the kind under discussion increase as the number of parameters to be estimated increases. In many cases, as the number of parameters to be estimated approaches infinity, the estimation errors become equal to the observation errors so that all smoothing is lost.

For example, suppose

$$S(t) = x\ (t) + \varepsilon(t) \tag{50}$$

where x(t) is some function of time, the signal, to be estimated, and $\varepsilon(t)$ is the noise. If we allow x(t) to be represented by a countable infinity of parameters without a-priori statistics, and then apply generalized least-squares smoothing, the resulting estimates are

$$\hat{x}(t) = S(t) \tag{51}$$

and the estimation error is just $\varepsilon(t)$.

Ordinarily, one gets smoothing by fitting x(t) by a set of functions depending on a small number of parameters, such as low-order polynomials or trigonometric series. This reduces the estimation errors due to observation noise, but if the actual functions x(t) do not belong precisely to the set of functions used in the fitting procedure, another kind of error is introduced which is sometimes called "bias error" (although it has nothing to do with biases in the observation errors).

Usually, the procedure is "optimized" by choosing the number of parameters, e.g., the order of the polynomial or the number of terms in the trigonometric series, so that the sum of the "bias" errors and the errors due to observation noise is minimized. This "optimization" is facilitated if one has some sort of a-priori knowledge as to how closely the functions x(t) which actually characterize the observations can be approximated by functions belonging to the set used to fit the observations.

According to the viewpoint adopted here, smoothing can be retained even though x(t) continues to be represented by an infinite (countable) parameter set, provided these parameters are given a joint a-priori statistical distribution. This is equivalent to adding an infinite set of virtual observations which is sufficient to retain full smoothing even with an infinite number of parameters.

Of course, another way of looking at it would be that this is equivalent to regarding x(t) as a random process, and is in fact just another way of interpreting the standard linear optimum filtering in which the signal as well as the noise is regarded as a random process.

This equivalence will be made explicit in a moment. However, it would be well to mention, at this point, that examples can be found of problems in which there are an infinite number of unknown parameters, but in which smoothing can be obtained without associating a-priori statistics with any of the unknown parameters. An example of this sort is given in the appendix.

To make the above-described interpretation of linear optimum filtering explicit, suppose

$$S(t) = x(t) + \varepsilon(t), \ \tau_1 \leqq t \leqq \tau_2 \tag{52}$$

where $x(t)$ and $\varepsilon(t)$ are sample functions of random processes $\{x(t)\}$, $\{\varepsilon(t)\}$. It is assumed that one knows a-priori that $\{x(t)\}$ is defined and continuous in the mean over an interval (T_1, T_2) containing (τ_1, τ_2), with zero mean and covariance function $\Phi_x(s, t)$; while $\{\varepsilon(t)\}$ is defined and continuous in the mean over (τ_1, τ_2) with zero mean and covariance function $\Phi_\varepsilon(s, t)$. $\{x(t)\}$ and $\{\varepsilon(t)\}$ are assumed to be statistically uncorrelated.

Now, let $\hat{x}(t)$ be the estimate of $x(t)$ obtained from standard linear least squares theory for $t \epsilon (T_1, T_2)$. (If $t \epsilon (\tau_1, \tau_2)$ we have a true filtering or interpolation problem, otherwise an extrapolation problem.)

Now, consider the following equivalent problem. Suppose the actual observations are $S(t)$, $\tau_1 \leqq t \leqq \tau_2$. Also suppose the virtual observations are defined as follows:

$$\bar{x}(t) = \text{"observed value" of } S^{(e)}(t) = 0, \quad T_1 \leqq t \leqq T_2 \tag{53}$$

(since our assumption is that $\{x(t)\}$ is a zero mean process).

$$\text{Virtual observation error} = \{\varepsilon^*(t)\}, \quad T_1 \leqq t \leqq T_2 \tag{54}$$

where $\{\varepsilon^*(t)\}$ is a zero mean random process with covariance function $\Phi_x(s, t)$. (The actual observation error is the same as before.)

Let $\hat{x}^*(t)$ be the generalized least squares estimate obtained for this equivalent version of the problem, $t \epsilon (T_1, T_2)$. Then,

$$\hat{x}^*(t) = \hat{x}(t) \tag{55}$$

This is actually a consequence of the results previously proved; however, a direct verification is possible. This can most simply be obtained by considering the discrete-time case in which t is restricted to discrete values $\{t_\mu\}$. The estimates for the continuous time parameter can be obtained by a limiting process from the discrete-time results.[2, 14]

The estimation process for the equivalent least squares smoothing technique takes the following form: let

$$Q(x) = \sum_{\mu,\upsilon}{}' \eta_{\mu\upsilon}\left[S(t_\mu) - x(t_\mu)\right]\left[S(t_\upsilon) - x(t_\upsilon)\right] \qquad (56)$$

$$+ \sum_{\mu,\upsilon} \zeta_{\mu\upsilon}\, x(t_\mu)\, x(t_\upsilon)$$

where
$$\eta = \Phi_{\varepsilon}^{-1} \qquad (57)$$

$$\zeta = \Phi_{x}^{-1}$$

The parameters to be estimated are $x_\mu = x(t_\mu)$, $t_\mu\, \varepsilon(T_1, T_2)$. The second sum in Eq. (56) is extended over the whole interval (T_1, T_2), while the first sum is extended over only those t_μ in (τ_1, τ_2), as indicated by the prime symbol. The parameter estimates $\hat{x}^*(t_\mu)$ are obtained by minimizing Q(x) with respect to $x(t_\mu)$.

The fact that Eq. (55) holds, where $\hat{x}(t_\mu)$ are the estimates resulting from standard linear least squares filtering theory, can be verified directly by minimizing Q(x) with respect to $x(t_\mu)$ by setting the partials $(\partial Q \mid \partial x_\mu) = 0$, and comparing the results with the standard formulas for $\hat{x}(t_\mu)$, as for example given in Ref. 14.

As a specific example, suppose $(T_1, T_2) = (\tau_1, \tau_2)$. The estimates $\hat{x}(t_\mu)$ resulting from minimization of Q in Eq. (56) are given (in vector notation) by

$$\hat{x} = (\eta + \zeta)^{-1} \eta S \tag{58}$$

which is easily shown to be identical to the more familiar form[14]

$$\hat{x} = \zeta^{-1}(\eta^{-1} + \zeta^{-1})^{-1} S \tag{58a}$$

Incidentally, insofar as concerns estimation of any particular value $x(t_o)$, $t_o \in (T_1, T_2)$, the estimate $\hat{x}^*(t_o)$ requires only that the interval over which the random process $\{x(t)\}$ is defined contain (τ_1, τ_2) and the point t_o; $\hat{x}^*(t_o)$ will be independent of whether $\{x(t)\}$ is actually defined over the full interval (T_1, T_2) if the latter is larger than (τ_1, τ_2). This is also, of course, true for $\hat{x}(t_o)$.

Another result which is a direct consequence of the results stated at the end of Section II. 2 is as follows:

Let

$$S(t) = \sum_{i=1}^{n} x^{(i)}(t) \tag{59}$$

where $x^{(i)}(t)$ are sample functions from random processes $\{x^{(i)}(t)\}$ which have zero means, covariance functions $\Phi^{(i)}(s, t)$ and are mutually uncorrelated.

Suppose the indices $i = 1, \ldots, n$ are divided into two sets, I and I', and that the processes $\{x^{(i)}(t)\}$ for $i \in I$ are regarded as "signals", while the "noise" is given by

$$\mathcal{E}(t) = \sum_{i \varepsilon I'} x^{(i)}(t) \qquad (60)$$

Then, the estimate $\hat{x}^{(i)}(t)$ for a particular index $i = i_o \varepsilon I$ is independent of $I - i_o$. That is, so long as $i_o \varepsilon I$, the estimate $\hat{x}^{(i)}(t)$ for $i = i_o$ is the same regardless of which of the remaining processes $x^{(i)}(t)$, $i \neq i_o$, are considered as signals to be jointly estimated, and which are lumped into the noise.

We can even go to the extreme of considering all of the processes, $i = 1, \ldots, n$, as signals to be jointly estimated. In the equivalent generalized least squares formulation, the "actual" observations would then be considered to be error-free and the only errors would be associated with the "virtual" observations.

III. Recursive Solutions of Signal Estimation Problems

III. 1. Preliminary Discussion

Suppose one has a signal vector x = $(x_1, \ldots, x_m)$, which may be a function of time; observational data which depend on the values of the vector x; and additive observation errors. Recursive methods for producing the generalized least squares estimate of $\{x_i\}$ at any time t have been studied by Swerling,(9, 10) Kalman,(11, 12) Bucy,(12) Blum,(13) and others. These solutions have the feature that optimum estimates based on previous data are combined with additional observational data in an optimum way to produce new optimum estimates.

Swerling(9, 10) treated initially the case (either linear or non-linear) where the vector x is constant; then the modified case where x may depend on time but where the variation of x with time has known functional form; and finally the case where the variation of x with time has a component of unknown functional form, but without associating a-priori statistics with the unknown components of the time variation of x.

Kalman(11,12) and Bucy(12) treat the linear case, and also give the extension to the case where x is regarded as a random process, with essentially the assumption that both the signal x and the observation noise are projections of vector Markov processes.

Blum(13) generalizes these recursive methods, in a manner somewhat different from the other papers mentioned, to cases where the observation errors are correlated.

It is the purpose of this section to exhibit recursive solutions which yield the linear optimum estimates for the case where x is a random process, with very few restrictive assumptions on the statistics of x or of the noise process. Essentially, the signal processes are assumed to be continuous in the mean; the noise process is assumed to consist of a component which is continuous in the mean and a white noise component; and that is all. Section III. 2 treats the linear case with additive noise; and Section III. 3 gives the "first order" treatment of the non-linear case, where some of the noise may be non-additive.

The method of approach is as follows: all problems of this type are reduced to an equivalent problem in which

(a) There is a (possibly) infinite set of parameters to be estimated

(b) The parameters to be estimated are regarded as constants (independent of time)

(c) A-priori statistics are associated with the parameters to be estimated and are represented in the generalized least squares procedure in the form of equivalent virtual observations.

(d) The observation errors are regarded as uncorrelated, i.e., have covariance function $\Phi(s, t) = R(t)\ \delta(s - t)$. This is accomplished by regarding everything except the "white noise" component of the observation error as represented by parameters to be estimated.

When the problem has been reduced to the form described by (a) - (d), formulas of Swerling$^{(9)}$ can be applied directly. The requisite formulas will be reproduced here, in the form applicable to the discrete-time case. The result for a continuous time parameter is then obtained by a limiting process.

Suppose we have observed data given by Eq. (12) above. Suppose the matrix (η_{μ},) can be broken up into blocks. The fundamental result of Swerling[(9)] is that a recursive procedure can be set up, in which the observational data corresponding to each block of η is treated as a separate stage; at each stage, say the s^{th}, a generalized least squares smoothing of the observation data in the s^{th} stage, together with the estimates based on all previous stages, is defined. The basic result is that this sequence of generalized least squares estimates can be defined in such a way that the resulting estimates (say, after the s^{th} stage) are, tc first order, identical with those resulting from the non-recursive smoothing of all "s" stages using the original matrix η. In the linear case, the qualifying phrase "to first order" can be dropped. The specific form of the necessary recursive sequence is exhibited.

The results assume a particularly simple form if the original smoothing matrix η is diagonal (or at least is diagonal after some point). In this case, each observation S_{μ} can be considered a separate stage. We will refer to this as introducing the observations one-by-one.

Before exhibiting the formulas necessary for the ensuing application, a few comments are in order about the interpretation of these stagewise or recursive procedures. The most important comment is that Swerling's basic result[(9)] need not be interpreted statistically, that is, it can be stated and proved without recourse to statistical notions; it holds regardless of the statistics of the observation error, and in particular, of whether the basic smoothing matrix η is or is not the inverse covariance matrix of the observation errors. Consequently, even if the errors are correlated, one is still at liberty to use a diagonal η and thus to use one-by-one recursive methods, although the result may not be statistically optimum.

Although the basic result need not be interpreted statistically, all the usual statistical consequences can be derived from it when specific statistics are associated with the observation errors. For example, if conditions A and B of Section II hold, the result of the non-recursive method are ML estimates, and consequently the result of the recursive method are to first order the ML estimates.

If the original weighting matrix η is not the inverse covariance matrix of the observation errors, then the accuracy of the resulting estimates will be degraded (the estimates will not be statistically optimum); the amount of accuracy degradation can be computed[(2)] and may be an acceptable price to pay, for example, for the computational convenience of using a diagonal matrix η.

Blum[(13)] considers recursive estimation procedures that can be applied without loss of statistical optimality to certain cases where the observation errors are correlated and where, in fact, there is no way to break up their covariance matrix into blocks. His approach is to assume that the observation errors satisfy a non-homogeneous difference equation of order, say, k; his recursive procedure then involves the previous k + 1 estimates instead of just the last estimate.

The approach to be followed here (summarized above in statements a - d) is applicable to either correlated or uncorrelated observation errors, and results in all cases in statistically optimum estimates (under the appropriate conditions A' and B'). By employing step (d), regarding the correlated part of the errors as parameters to be estimated, the problem is reduced to one in which a diagonal η may be used without loss of statistical optimality. Also, by reducing the entire problem to an equivalent one where the parameters all appear as non-stochastic, it is possible to dispense with any assumption that either signal or observation noise are projections of Markov processes. Alsc, no matrix inversions are required at any stage of the procedure. The necessary equations are equivalent to Eqs. (23), (25), (46), (47), (48) of Ref. 9.

Thus, suppose the matrix $(\eta_{\mu\upsilon})$ is diagonal:

$$(\eta_{\mu\upsilon}) = \eta_\mu \, \delta_{\mu\upsilon} \tag{61}$$

Let $x = (x_1, \ldots, x_m)$ where x_i are constants. Denote by $\hat{x}_i(s)$ the estimate of x_i based on the first s observations. (We will also assume that there is some initial estimate $\hat{x}_i(o)$ but need not specify at this point just how this is obtained.)

Then, the one-by-one recursive procedure is defined as follows, where $\hat{x}(s) = \left[\hat{x}_1(s), \ldots, \hat{x}_m(s)\right]$:

$$\hat{x}(s) - \hat{x}(s-1) = D^{(s)}\left[\hat{x}(s-1)\right]_\rho *\left[\hat{x}(s-1)\right] \tag{62}$$

$$D_{ij}^{(s)}\left[x\right] = D_{ij}^{(s-1)}(x) - d_i^{(s)}(x)\, d_j^{(s)}(x) \tag{63}$$

$$d_i^{(s)}(x) = \left\{\sqrt{\eta_s} \sum_{k=1}^{m} \frac{\partial f_s(x, \tau_s)}{\partial x_k} D_{ik}^{(s-1)}(x)\right\} \tag{64}$$

$$\times \left\{1 + \eta_s \sum_{j,k=1}^{m} D_{jk}^{(s-1)} \frac{\partial f_s(x,\tau_s)}{\partial x_j} \frac{\partial f_s(x,\tau_s)}{\partial x_k}\right\}^{-\frac{1}{2}}$$

$$\rho_i^\star\left[\hat{x}(s-1)\right] = \eta_s \frac{\partial f_s\left[\hat{x}(s-1), \tau_s\right]}{\partial x_i} \left\{S(t_s) - f_s\left[\hat{x}(s-1), \tau_s\right]\right\} \tag{65}$$

In the above, $\hat{x}$, $\rho^\star$, and d are m-component vectors; $D^{(s)}$ is an m × m matrix; τ_s is the time of the s^{th} observation.

The initial values of $\hat{x}$ and D, i.e., $\hat{x}(o)$ and $D^{(o)}$, have not been defined; but this will be done when the intended application is made.

In the linear case, the matrices $D^{(s)}$ and the partials $\partial f^{(s)}/\partial x_i$ are independent of the values of x or $\hat{x}(s)$.

III. 2. Application to the Linear Case

We will assume that the observational data up to any time instant τ is

$$S(t) = \sum_{i=1}^{n-1} u^{(i)}(t) + \varepsilon^\star(t), \quad \tau_1 \leqq t \leqq \tau \tag{66}$$

where $u^{(i)}(t)$ is a sample function of a random process $\left\{u^{(i)}(t)\right\}$. The processes $\left\{u^{(i)}(t)\right\}$ will be interpreted as signal processes and $\left\{\varepsilon^{*}(t)\right\}$ as a noise process.

It will become obvious that the same technique may be applied to more general situations in which there is not one, but say, J received signals, each containing a different linear combination of the random processes $\{u^{(i)}(t)\}$:

$$S_j(t) = \sum_{i=1}^{n-1} a_{ij}(t)\ u^{(i)}(t) + \varepsilon_j^{*}(t) \tag{66a}$$

with $j = 1, \ldots, J$. The method of doing this will be outlined below.

The following assumptions will be made about these random processes:

There is some basic interval (T_1, T_2) containing (τ_1, τ) within which it is known a-priori that:

(a) $\left\{u^{(i)}(t)\right\}$, $i = 1, \ldots, n - 1$, are random processes which are mutually uncorrelated; have zero means; and are continuous in the mean with covariance function $\phi^{(i)}(t, t')$.

(b) $\left\{\varepsilon^{*}(t)\right\}$ is a random process which is uncorrelated with all the processes $\left\{u^{(i)}(t)\right\}$, and which can be written

$$\left\{\varepsilon^{*}(t)\right\} = \left\{\gamma(t)\right\} + \left\{\varepsilon(t)\right\} \tag{67}$$

where $\{\gamma(t)\}$ and $\{\varepsilon(t)\}$ are mutually uncorrelated, zero mean, and:

(i) $\{\gamma(t)\}$ is continuous in the mean with covariance function $\Phi^{(n)}(t, t')$.

(ii) $\{\varepsilon(t)\}$ is a "generalized white noise" process.

The statement about $\{\varepsilon(t)\}$ can be interpreted in various alternative ways:

1. $\{\varepsilon(t)\}$ has covariance function $R(t)\ \delta(t - t')$

2. If we define (more rigorously) another random process by

$$\beta(\Delta t, t) = \int_{t}^{t+\Delta t} \varepsilon(\tau)\ d\tau \tag{68}$$

then in the neighborhood of t, $\beta(\Delta t, t)$ has covariance function

$$E\left[\beta(\Delta t, t)\ \beta(\Delta t', t)\right] = R(t) \min(\Delta t, \Delta t') \tag{69}$$

3. If $R(t) =$ constant, then $\{\varepsilon(t)\}$ has one-sided spectral density

$$N_o = 2\ R \tag{70}$$

It will be assumed that for $t\ \epsilon (T_1, T_2)$, $R(t)$ is positive and bounded away from zero. On the other hand, the non-white noise component may vanish.*

*Recently, the case has been investigated where the white noise component vanishes(15).

Now, if the non-white noise component $\nu(t)$ does not vanish, we will define

$$\left\{u^{(n)}(t)\right\} = \left\{\nu(t)\right\} \tag{71}$$

and write

$$S(t) = \sum_{i=1}^{n} u^{(i)}(t) + \varepsilon(t) \tag{72}$$

(If $\left\{u^{(n)}(t)\right\}$ vanishes, the sum in Eq. (72) simply extends to n - 1 instead of n.) Incidentally, the assumption that the processes $\left\{u^{(i)}(t)\right\}$ are mutually uncorrelated is not really essential; it will be made clear how the same technique could be applied in the absence of this assumption.

Now, it will be assumed that the quantities to be estimated are $u^{(i)}(t)$, $t \in (T_1, T_2)$, $i = 1, \ldots, n$. (Actually one is really only interested in $i = 1, \ldots, n - 1$, but the technique calls for treating the non-white noise component as if it were an n^{th} signal component to be estimated.)

Let

$\hat{u}^{(i)}(t, \tau)$ = linear optimum estimate of $u^{(i)}(t)$ based on the observational data from τ_1 to τ, for any $t \in (T_1, T_2)$. (73)

The actual observed data are given by Eq. (72). The virtual observations are given by

$$\bar{u}^{(i)}(t) = 0, \; i = 1, \ldots, n, \quad t \, \epsilon (T_1, T_2) \qquad (74)$$

and the virtual observation errors are random processes having the same statistics as $\left\{u^{(i)}(t)\right\}$.

The solution for $\hat{u}^{(i)}(t, \tau)$ will be obtained by first treating the discrete-time case and later passing to the limit.

Thus, let $\left\{t_\mu\right\}$ be a set of equally spaced time points in (T_1, T_2):

$$t_{\mu+1} - t_\mu = \Delta t \qquad (75)$$

$$t_1 = T_1$$

Let $\left\{\tau_s\right\}$ be a set of equally spaced time points in (τ_1, τ):

$$\tau_{s+1} - \tau_s = \Delta t \qquad (76)$$

The parameters to be estimated are $u^{(i)}(t_\mu)$. The actual observations are given by Eq. (72) with t restricted to the points $\left\{\tau_s\right\}$. The virtual observations are given by Eq. (74) with t restricted to the points $\left\{t_\mu\right\}$. It is assumed that the points $\left\{\tau_s\right\}$ are a subset of $\left\{t_\mu\right\}$.

It is assumed that the "zeroth" stage of the recursive procedure is based only on the virtual observations, that is, on the a-priori statistics of $\left\{u^{(i)}(t)\right\}$. Thus, let

$$\hat{u}^{(i)}(t_\mu, s) = \text{estimate of } u^{(i)}(t_\mu) \text{ based on the first s actual observations and the a-priori statistics} \qquad (77)$$

$$\hat{u}^{(i)}(t_\mu, 0) = \bar{u}^{(i)}(t_\mu, 0) \equiv 0 \qquad (78)$$

The various covariance functions $\Phi^{(i)}(t, t')$ become covariance matrices $\Phi^{(i)}(t_\mu, t_\nu)$ in the obvious manner. The only point that needs discussion is the manner in which the white noise process $\left\{\varepsilon(t)\right\}$ is to be represented in the discrete-time case. The correct representation is

$$\Phi_\varepsilon(t_\mu, t_\nu) = \delta_{\mu\nu} \frac{R(t_\mu)}{\Delta t} \qquad (79)$$

where Φ_ε is the covariance matrix of $\left\{\varepsilon(t_\mu)\right\}$.

The solution for the discrete-time case then consists of applying Eq. (62) - (65). This is straightforward, the main difficulty being in keeping all the indices straight. Here, the parameter vector $x = (x_1, \ldots, x_m)$ has components $u^{(i)}(t_\mu)$ arranged in some order. Thus, $m >> n$ in general.

The result is as follows. It should be noted that the indices i, j, k, ℓ in the following run over $(1, \ldots, n)$; thus, they do not have quite the same meaning as in Eqs. (62) - (65). In fact, the indices i appearing in Eqs. (62) - (65) correspond to pairs of indices (i, μ) in the following. Also, the functions C below arise from the quantities labeled D in Eqs. (62) - (65). It is best to regard the following equations as self-contained; Eqs. (62) - (65) were merely reproduced to indicate the method of derivation, and the notation of Eqs. (62) - (65) is not necessarily completely consistent with that of the following.

$$\hat{u}^{(i)}(t_\mu, s) - \hat{u}^{(i)}(t_\mu, s - 1) \tag{80}$$

$$= \frac{\Delta t}{R(\tau_s)} \left[S(\tau_s) - \sum_{j=1}^{n} \hat{u}^{(j)}(\tau_s, s - 1)\right] \sum_{k=1}^{n} C_{ik}(t_\mu, \tau_s, s)$$

$$\hat{u}^{(i)}(t_\mu, 0) = 0, \quad \text{all } i, \mu \tag{81}$$

$$C_{ij}(t_\mu, t_\upsilon, s) = C_{ij}(t_\mu, t_\upsilon, s - 1) \tag{82}$$

$$- d_i(t_\mu, s)\, d_j(t_\upsilon, s)$$

$$d_i(t_\mu, s) = \left[\frac{\Delta t}{R(\tau_s)}\right]^{\frac{1}{2}} \sum_{k=1}^{n} C_{ik}(t_\mu, \tau_s, s - 1) \tag{83}$$

$$\times \left\{ 1 + \left[\frac{\Delta t}{R(\tau_s)}\right] \sum_{j,k=1}^{n} C_{jk}(\tau_s, \tau_s, s - 1)\right\}^{-\frac{1}{2}}$$

$$C_{ij}(t_\mu, t_\nu, 0) = \delta_{ij} \; \Phi^{(i)}(t_\mu, t_\nu) \tag{84}$$

The case where the random processes $\left\{u^{(i)}(t)\right\}$ are mutually correlated is obtained simply by changing Eq. (84) to

$$C_{ij}(t_\mu, t_\nu, 0) = \Phi^{(i,j)}(t_\mu, t_\nu) \tag{84 a}$$

where, denoting a-priori expected value by E( ),

$$\Phi^{(i,j)}(t_\mu, t_\nu) = E\left[u^{(i)}(t_\mu)\; u^{(j)}(t_\nu)\right] \tag{85}$$

We will now go to the limit by assuming that $\Delta t \to 0$. Then we get

$$\frac{\partial}{\partial \tau}\left[\hat{u}^{(i)}(t, \tau)\right] = \frac{1}{R(\tau)}\left[S(\tau) - \sum_{j=1}^{n} \hat{u}^{(j)}(\tau, \tau)\right] \times \sum_{k=1}^{n} C_{ik}(t, \tau, \tau) \tag{86}$$

$$\hat{u}^{(i)}(t, \tau_1) = 0, \text{ all } i \text{ and all } t \in (T_1, T_2).$$

$$\frac{\partial}{\partial \tau}\left[C_{ij}(t, t', \tau)\right] = -\frac{1}{R(\tau)}\left\{\sum_{k,\ell=1}^{n} C_{ik}(t, \tau, \tau)\, C_{j\ell}(t', \tau, \tau)\right\} \tag{87}$$

$$C_{ij}(t, t', \tau_1) = \delta_{ij}\ \Phi^{(i)}(t, t') \tag{88}$$

for all i, j and all t and t′ in (T_1, T_2).

More generally, if $\left\{u^{(i)}(t)\right\}$ are mutually correlated,

$$C_{ij}(t, t', \tau_1) = \Phi^{(i,j)}(t, t') \tag{88 a}$$

In the above, $\hat{u}^{(i)}(t, \tau)$ is the estimate of $u^{(i)}(t)$, $t\ \epsilon (T_1, T_2)$, based on the observational data from τ_1 to τ. The indices i, j, k, ℓ as previously stated run from one to n.*

In the linear case which has been treated in this section, the estimates $\hat{u}^{(i)}(t, \tau)$ are precisely the minimum mean square error estimates of $u^{(i)}(t)$, based on actual observations up to τ.

The functions $C_{ij}(t, t', \tau)$ have the following interpretation (assuming the equivalent of conditions A′ and B′ of Section II apply):

$$C_{ij}(t, t', \tau) = E\left[\hat{u}^{(i)}(t, \tau) - u^{(i)}(t)\right] \tag{89}$$

$$\times \left[\hat{u}^{(j)}(t', \tau) - u^{(j)}(t')\right] \quad .$$

*It should also be mentioned that, since $S(\tau)$ is assumed to contain a white noise process, from the point of view of mathematical rigor Eq. (86) should actually be written with both sides integrated with respect to τ.

The expectation in Eq. (89) can be interpreted in either of two ways: the conventional way, regarding $u^{(i)}(t)$ and $u^{(j)}(t')$ as random variables and $\hat{u}^{(i)}$, $\hat{u}^{(j)}$ as random variables defined on the sample space of the "actual" observations only; or alternately, regarding $u^{(i)}(t)$ and $u^{(j)}(t')$ as constants, and $\hat{u}^{(i)}$, $\hat{u}^{(j)}$ as random variables defined on the sample space of the "actual" and the "virtual" observations.

It is useful to have the discrete-time formulas, Eqs. (80) - (84), since, in the first place, in many applications the observed data will be at discrete times; and second, even in the continuous time case, the solutions to Eqs. (86) - (88) would generally be built up from a difference equation approximation. Since the possible difference equation approximations are non-unique, Eqs. (80) - (84) indicate the best one (cf. especially Eq. (83), of which the term in braces disappears when $\Delta t \rightarrow 0$).

As an extension of the foregoing, suppose there are J observed processes, as in Eq. (66a).

Suppose the noise processes can, as before, be broken up into non-white and white components. The non-white components will be considered as J additional "signal" processes, and need not be statistically uncorrelated.

In order to preserve the feature of one-by-one addition of observations, it is necessary to assume that the J white noise components are mutually uncorrelated. If this is not true originally, one can transform the problem so that it is true as follows.

At any time t, assume there is a non-singular $J \times J$ matrix $M(t)$ which, applied to the J-vector of white noise components at time t, will transform them into a vector of uncorrelated components.

Then, simply consider the problem where the observations consist of the set of processes $\{S_j'(t)\}$ resulting from applying $M(t)$ to $\{S_j(t)\}$. We would have

$$S_j'(t) = \sum_{i=1}^{n-1} a'_{ij}(t)\, u^{(i)}(t) + \mathcal{E}_j'(t) \tag{66b}$$

but the white noise components would be mutually uncorrelated. (The prime symbol does not indicate differentiation.)

The most general form of the solution would be as follows. Let $\hat{u}^{(i)}(t \mid \tau^{(1)}, \ldots, \tau^{(J)})$ be the estimate of $u^{(i)}(t)$ based on observing the received processes $\{S_j(t)\}$ from $\tau_1^{(j)}$ to $\tau^{(j)}$, $j = 1, \ldots, J$. Also, i now ranges from 1 to $n + J - 1$. The solution gives the partial derivatives of $\hat{u}^{(i)}$ and C_{ik} with respect to the variables $\tau^{(j)}$, $j = 1, \ldots, J$.

As an example of the analysis in this section, suppose we wish to apply this to estimate the state of a dynamical system with a stochastic driving function. Specifically, suppose

$$\Lambda u = \eta \tag{90}$$

where u and η are vector random processes and Λ is a linear operaιor. It is unnecessary to assume either that η is a process of independent increments or that Λ is a differential operator. Λ and η are subject only to the conditions that

a. the a-priori covariance function of u be uniquely determined by that of η and the form of Λ; and

b. in the continuous-time case, that u be continuous in the mean.

The procedure is to solve for the covariance function of u and then apply the foregoing analysis.

Also, the partial differential equations above can, in the Markov case, easily be put into the form of total differential equations with respect to the latest observation time τ, in cases where the estimation time coincides with τ or is related to it by a fixed lead or lag.

III. 3. Non-Linear and Non-additive Application

Suppose the actual observation data is given by

$$S(t) = f\left[t, u^{(1)}(t), \ldots, u^{(n)}(t)\right] + \varepsilon(t), \quad \tau_1 \leqq t \leqq \tau \tag{91}$$

where $\left\{u^{(i)}(t)\right\}$ are mutually uncorrelated continuous-in-the-mean random process over (T_1, T_2), with covariance functions $\Phi^{(i)}(t, t')$ and means zero; $\left\{\varepsilon(t)\right\}$ is a generalized white noise process with covariance function $R(t)\,\delta(t - t')$, uncorrelated with the processes $\left\{u^{(i)}(t)\right\}$, and with zero mean.

It is assumed that some of the processes $\left\{u^{(i)}(t)\right\}$ may be considered "signal" and others "noise" from the point of view of any particular application; our technique calls, however, for all of them to be jointly estimated. Also, some of them may be additive and others not.

The recursive generalized least squares estimates $\hat{u}^{(i)}(t, \tau)$ can also be derived for this case by means of Eqs. (62) - (65). To make the first order approximations valid for the non-linear case we must now assume that at any time $\tau \geqq \tau_1$ there exist estimates of $u^{(i)}(t)$ which have small errors. We will assume that such estimates are, in fact, given by $\hat{u}^{(i)}(t, \tau)$. In the discrete-time case this becomes $\hat{u}^{(i)}(t_\mu, s)$.

In short, we will assume in the discrete-time case that $\hat{u}^{(i)}(t_\mu, s) - u^{(i)}(t_\mu)$ is sufficiently small so that

$$f\left[t, u^{(1)}(t), \ldots, u^{(n)}(t)\right] \tag{92}$$

$$= f\left[t, \hat{u}^{(1)}(t, s), \ldots, \hat{u}^{(n)}(t, s)\right]$$

$$+ \sum_{i=1}^{n} \frac{\partial f}{\partial u_i} \left[t, \hat{u}^{(1)}(t, s), \ldots, \hat{u}^{(n)}(t, s)\right]\left[u^{(i)}(t) - \hat{u}^{(i)}(t, s)\right]$$

+ negligible remainder.

In the continuous-time case, simply replace $\hat{u}^{(i)}(t, s)$ by $\hat{u}^{(i)}(t, \tau)$.

In the non-linear, non-additive case, the resulting generalized least squares estimates can no longer necessarily be stated to be minimum mean square error estimates. However, they can still be said to be asymptotically minimum mean square error if Eq. (92) holds.

The main difference between this and the linear, additive case previously discussed is that the functions C_{ij} now depend on estimates of $u^{(k)}(t)$, $k = 1, \ldots, n$. The necessary modification of Eqs. (80) - (84) or Eqs. (86) - (89) are as follows:

In the discrete-time case, writing as usual $u = (u^{(1)}, \ldots, u^{(n)})$,

$$\hat{u}^{(i)}(t_\mu, s) - \hat{u}^{(i)}(t_\mu, s - 1) \tag{93}$$

$$= \frac{\Delta t}{R(\tau_s)} \left\{ S(\tau_s) - f\left[\tau_s, \hat{u}(\tau_s, s - 1)\right]\right\}$$

$$\times \sum_{k=1}^{n} C_{ik}(t_\mu, \tau_s, s) \frac{\partial f}{\partial u_k} \left[\tau_s, \hat{u}(\tau_s, s - 1)\right]$$

$$\hat{u}^{(i)}(t_\mu, o) = 0, \text{ all } i \text{ and } \mu$$

$$C_{ij}(t_\mu, t_\nu, s) = C_{ij}(t_\mu, t_\nu, s - 1) \tag{94}$$

$$- d_i(t_\mu, s)\, d_j(t_\nu, s)$$

$$d_i(t_\mu,\ s) \ = \ \left[\frac{\Delta\ t}{R(\tau_s)}\right]^{\frac{1}{2}} \sum_{k=1}^{n} C_{ik}(t_\mu,\ \tau_s,\ s\ -\ 1)\ \frac{\partial\ f}{\partial\ u_k}\left[\tau_s,\ \hat{u}(\tau_s,\ s\ -\ 1)\right] \qquad (95)$$

$$\times\ \left\{1 + \left[\frac{\Delta\ t}{R(\tau_s)}\right] \sum_{j,k=1}^{n} C_{jk}(\tau_s,\ \tau_s,\ s\ -\ 1)\ \frac{\partial\ f}{\partial\ u_j}\left[\tau_s,\ \hat{u}(\tau_s,\ s\ -\ 1)\right] \frac{\partial\ f}{\partial\ \ _k}\left\lfloor\tau_s,\ \hat{u}(\tau_s, s-1)\right]\right\}^{-\frac{1}{2}}$$

$$C_{ij}(t_\mu,\ t_\upsilon,\ o) \ = \ \delta_{ij}\ \Phi^{(i)}(t_\mu,\ t_\upsilon) \qquad (96)$$

When $\Delta\ t \rightarrow o$, we get

$$\frac{\partial}{\partial\ \tau}\left[\hat{u}^{(i)}(t,\ \tau)\right] \ = \ \frac{1}{R(\tau)}\ \left\{S(\tau)\ -\ f\left[\tau,\ \hat{u}(\tau,\ \tau)\right]\right\} \qquad (97)$$

$$\times \sum_{k=1}^{n} C_{ik}(t,\ \tau,\ \tau)\ \frac{\partial\ f}{\partial\ u_k}\ \left[\tau,\ \hat{u}(\tau,\ \tau)\right]$$

$$\hat{u}^{(i)}(t,\ \tau_1) \ = \ 0, \quad \text{all } t\ \epsilon(T_1,\ T_2) \qquad (98)$$

$$\frac{\partial}{\partial\ \tau}\ \left[\ C_{ij}(t,\ t',\ \tau)\right] \qquad (99)$$

$$= -\ \frac{1}{R(\tau)} \sum_{k,\ell=1}^{n} C_{ik}(t,\tau,\tau)\ C_{j\ell}(t',\tau,\tau)\ \frac{\partial\ f}{\partial\ u_k}\ \left[\tau,\ \hat{u}(\tau,\tau)\right] \frac{\partial\ f}{\partial\ u_\ell}\ \left[\tau,\hat{u}(\tau,\tau)\right]$$

$$C_{ij}(t,\ t',\ \tau_1) \ = \ \delta_{ij}\ \Phi^{(i)}(t,\ t') \qquad (100)$$

As before, if the assumption is dropped that $\left\{u^{(i)}(t)\right\}$ are mutually uncorrelated, then Eq. (98) and Eq. (100) are replaced by Eqs. (84 a) or (88 a) respectively.

IV. Further Problems

Several areas of further research are suggested by the foregoing. Two specific areas of useful investigation are, first, use of the recursive framework to treat problems calling for adaptive estimation methods, and second, investigation of techniques for obtaining exact (not merely asymptotic) minimum mean square error estimates in the non-linear case.

Adaptive Estimation Methods

As used here, an adaptive estimation problem refers to one in which the a-priori statistics of the observation errors, or the statistics of the signals if these are regardeu as random processes, are not known exactly.

The generalized least squares methods described above, especially in their recursive form, may provide a convenient framework for treating such problems, as has been recognized by many workers in this field.

For example, considering the problems treated in Section III, suppose that the covariance functions $\Phi^{(i)}(t, t')$ and the function R(t) are not known exactly. How would one modify the recursive, generalized least squares procedure to incorporate a feature whereby these covariance functions are estimated from the data and these estimates then incorporated in the procedure? One possible avenue of approach is as follows.

Suppose the estimation procedure is initiated simply by assuming some set of function $\Phi^{(i)}(t, t')$ and $R(t)$ to insert into the recursive equations. As previously discussed, the resulting recursive estimates are still identical (to first order) to some set of generalized least squares estimates. However, these estimates in effect are obtained by minimizing a function Q in which the matrix η is not the true inverse covariance matrix of the observation errors (actual and virtual).

However, this may still result in reasonably good estimates, even though they will not be the best possible. Moreover, the machinery exists[(2)] by which it is possible to compute the estimation error covariances as a function of the deviation between the true covariance matrices and the assumed ones.

Now, this procedure will result in estimates $\hat{u}^{(i)}(t, \tau)$ of the random variables $u^{(i)}(t)$. Also, it can be used to produce estimates $\hat{\varepsilon}(t, \tau)$ of the "white noise" component, since $\hat{\varepsilon}(t, \tau) = S(t) - f\left[t, \hat{u}(t, \tau)\right]$. (In the continuous time case, the estimate $\hat{\varepsilon}(t, \tau)$ would have to be interpreted as an estimate of some suitably smoothed version of the white noise, since, technically, the white noise has infinite variance at a single instant.)

Next, the estimates $\hat{u}^{(i)}(t, \tau)$ and $\hat{\varepsilon}(t, \tau)$ can be used to estimate the covariance functions $\Phi^{(i)}(t, t')$ and $R(t)$. This statement is clear in case the random process $\left\{u^{(i)}(t)\right\}$ and $\left\{\varepsilon(t)\right\}$ are stationary, in which case estimates of their covariance functions can be made from a single time sample.

If they are non-stationary, it is still possible to get estimates of the covariance functions from a single time sample, but in general these would be very poor. However, if these unknown processes can be considered to be simple transformations of stationary processes, such as integrals of stationary processes, then the covariance function can be more accurately estimated from a single time sample.

It still remains to specify how one would incorporate the resulting estimates of the covariance functions (or possibly some composite estimates, depending on both the covariance functions assumed a-priori and those estimated from the data) into the overall estimation procedure.

Since, in the recursive procedure, the functions $\Phi^{(i)}(t, t')$ enter into the procedure only via the initial condition equations (84), (88), (96), or (100), one approach would be, periodically, to go back and solve for the functions $C_{ij}(t, t', \tau)$ over again, using $\hat{\Phi}^{(i)}(t, t', \tau)$ and $\hat{R}(t, \tau)$ in place of the initially assumed $\Phi^{(i)}(t, t')$, $R(t)$ in these equations. Here, $\hat{\Phi}$ and $\hat{R}$ refer to covariance function estimates making use of data up to time τ. The resulting recomputed values of $C_{ij}(t, \tau, \tau)$ would then be used from that point on in Eqs. (80), (86), (93), or (97). In part, this would detract from the recursive feature, since it would involve re-solving for $C_{ij}(t, t', \tau)$. However, it still preserves the recursive feature insofar as processing of new observational data is concerned (at least, this is true in the linear case), since it does not require any re-processing of the old observational data.

Insofar as the function R(t) is concerned, it would appear necessary to make some sort of assumption that R(t) is slowly varying.

Some specific problems to be investigated are

a) Computation of the degree to which the accuracy of the estimates $\hat{u}^{(i)}(t, \tau)$ is degraded if the initially assumed $\Phi^{(i)}(t, t')$ and R(t) differ from the true covariance functions of $\{u^{(i)}(t)\}$ and $\{\mathcal{E}(t)\}$. The machinery for this already exists.(2)

b) Computation of the degree to which the initially assumed functions can be improved on the basis of the observed data, whether and under what conditions the resulting estimates $\hat{\Phi}$ and $\hat{R}$ actually approach the true covariance functions, and if they do, how rapidly as τ increases.

c) Devising computationally convenient ways of incorporating improved estimates of the covariance functions into the procedure.

Exact Minimum Mean Square Error Estimates

Even if conditions A′ and B′ are satisfied, the generalized least squares estimation procedures do not give the precise minimum mean square error estimates for the non-linear case. However, it is known, at least formally, what the exact minimum mean square error estimates are. They are the estimates formed by finding the expected values of x_i with respect to the a-posteriori p d f of $(x_1, x_2, \ldots)$ based on the observational data.

Now, in some cases the mean of $p(x_1, x_2, \ldots \mid S)$ occurs for the same values of x_i as the maximum. In these cases, the MAP estimates are the minimum mean square error estimates. This is true in the linear case. However, in general it is not the case.

If one attempts to find the conditional means in the non-linear case, even for Gaussian additive noise, one quickly gets into analytical and computational problems of great difficulty.

Incidentally, the exact minimum variance unbiased estimates can be obtained from Barankin's method,[4] the application of which to stochastic processes is analytically tractable up to a point. However, in general, the Barankin minimum variance unbiased estimates are not the minimum mean square error estimates.

While the development of analytically or computationally tractable methods for finding the conditional means has been the subject of a considerable literature, further investigation of this problem would be very useful (and in fact forms the subject matter of Ref. 16.)

Other Applications

A number of applications can be thought of to problems in which there are various mixtures, all in the same problem, of parameters with which a-priori statistics are associated and those with which no a-priori statistics are associated; random processes involving mixtures of infinite sets of unknown parameters and additional finite sets of parameters; and mixtures of discrete-time and continuous-time observational data or other heterogeneous types of observational data. The foregoing results provide a systematic framework for treating large classes of these problems, including setting up recursive solutions.

The problem treated in Appendix A with its variations provides one set of examples. Numerous other specific applications can readily be thought of.

Appendix A. Estimation of Rotation Rate

This problem is introduced as an example of a problem in which a smoothed estimate of a parameter can be obtained, even though there are an infinite number of other unknown parameters, and even though no a-priori statistics are associated with the unknown parameters.

Suppose a radar target is rotating about a fixed axis. The returned signal amplitude is observed by the radar, and the object is to estimate the rotation rate ω. The radar cross-section of the target will be assumed to vary with viewing aspect. The estimate of ω is to be made only by observing the fluctuations in amplitude of the returned signal (and not by means of spreading of the doppler spectrum, for example).

It is assumed that nothing is known a-priori about the form of the radar cross-section vs. aspect (with two minor exceptions to be noted below).

The received signal is assumed to be

$$S(t) = \sigma(\omega\ t) + \varepsilon(t) \tag{A 1}$$

where $\{\varepsilon(t)\}$ is a white noise process with one-sided power spectral density Ψ.

If we write

$$\theta = \omega\ t \tag{A 2}$$

it is assumed that

(i) $\sigma(\theta)$ is periodic with period $2\pi/\omega$ but not with any smaller period.

(ii) $\theta\,\sigma'(\theta)$ is square integrable over $(o, 2\pi)$.

Other than this, no knowledge is assumed about σ.

Now, let $\{P_i(\theta)\}$ be any complete set of orthonormal functions in $L_2(o, 2\pi)$. Then we can write

$$\sigma(\theta) = \sum_{i=1}^{\infty} \alpha_i\, P_i(\theta) \tag{A 3}$$

Thus, the received signal is

$$S(t) = f(t, \omega, \alpha_1, \alpha_2, \ldots) + \mathcal{E}(t) \quad , o \leq t \leq T \tag{A 4}$$

where

$$f(t, \omega, \alpha_1, \alpha_2, \ldots) = \sum_{i=1}^{\infty} \alpha_i\, P_i(\omega\, t) \tag{A 5}$$

We now wish to find $E\left[\hat{\omega} - \omega\right]^2$ for the maximum likelihood estimate $\hat{\omega}$ of rotation rate. The approach will be to apply the formulas of Ref. 2 for the white noise case. This approach leads to the following.

$$E\left[\hat{\omega} - \omega\right]^2 \geq (B^{-1})_{oo} \tag{A 6}$$

with asymptotic equality,

where

$$B_{ij} = \frac{2}{\Psi} \int_0^T \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \, d\, t \qquad \text{(A 7)}$$

In Eq. (A 7), $i = 0, 1, 2, \ldots$

$j = 0, 1, 2, \ldots$

$x_0 = \omega$

$x_i = \alpha_i, \; i > 0$

Now, for convenience, suppose the observation time T is such that it is an integral number of periods:

$$T = \frac{2\pi\, N}{\omega} \qquad \text{(A 8)}$$

Then

$$B_{oo} = \frac{2}{\Psi\, \omega^3} \int_o^{2\pi\, N} \left[\theta\, \sigma'(\theta)\right]^2 d\, \theta \qquad \text{(A 9)}$$

$$B_{oi} = B_{io} = \frac{2}{\Psi\, \omega^2} \int_o^{2\pi\, N} \left[\theta\, \sigma'(\theta)\; P_i(\theta)\right] d\, \theta, \quad i > o \qquad \text{(A 10)}$$

$$B_{ij} = \frac{2\, N}{\Psi\, \omega}\, \delta_{ij} \quad , \qquad i,\, j > o \qquad \text{(A 11)}$$

One can now compute $(B^{-1})_{oo}$ by truncating the matrix B to an n × n matrix; finding $(B^{-1})_{oo}$ for this n^{th} order matrix, and then letting $n \rightarrow \infty$. The result is

$$(B^{-1})_{oo} = \left[B_{oo} - \sum_{i=1}^{\infty} B_{io}^{2}/B_{ii}\right]^{-1} \tag{A 12}$$

Using Eq. (A 9) - Eq. (A 11), this results in

$$E\left[\hat{\omega} - \omega\right]^{2} \tag{A 13}$$

$$\geq \frac{\Psi\,\omega^{3}}{2} \left\{ \int_{0}^{2\pi N} \left[\theta\,\sigma'(\theta)\right]^{2} d\,\theta - \sum_{i=1}^{\infty} \frac{1}{N} \left[\int_{0}^{2\pi N} \theta\,\sigma'(\theta)\,P_{i}(\theta)\,d\,\theta\right]^{2}\right\}^{-1}$$

or, using Eq. (A 8), an equivalent form is

$$E\left[\hat{\omega} - \omega\right]^{2} \tag{A 14}$$

$$\geq \frac{\Psi}{2}\left(\frac{2\pi}{T}\right)^{3} \left\{ \int_{0}^{2\pi} \left[\theta\,\sigma'(N\,\theta)\right]^{2} d\,\theta - \sum_{i=1}^{\infty} \left[\int_{0}^{2\pi} \theta\,\sigma'(N\,\theta)\,P_{i}(N\,\theta)\,d\,\theta\right]^{2}\right\}^{-1}$$

It might be noted from Eq. (A 13) or Eq. (A 14) that $E\left[\hat{\omega} - \omega\right]^{2} \rightarrow \infty$ for $N = 1$, i.e., for T less than or equal to one period. That this is what should happen is clear.

Now, suppose we compare Eq. (A 13) or Eq. (A 14) with the answer to be obtained when $\sigma(\theta)$ is a-priori known exactly. It might be noted that in practical cases, even if the nature of the scattering object were well known, a realistic assumption would have the absolute amplitude and the initial phase of rotation as unknown parameters. That is, a realistic model would have

$$S(t) = \alpha\,\sigma\left[\omega(t - t_o)\right] + \mathcal{E}(t) \tag{A 15}$$

with α, ω, and t_o unknown a-priori.

In the case where σ was assumed to be completely unknown, treated above, it was unnecessary to assume additional unknown parameters α and t_o since this was automatically taken care of by assuming that all the α_i in Eq. (A 3) were unknown.

The formulas for treating Eq. (A 15) are simple to apply.(2) However, for present purposes, let us make the somewhat unrealistic assumption that ω is the only unknown parameter. (The answer will then be a lower bound for the case where α, ω, and t_o are all unknown.) Thus, instead of Eq. (A 15), we will assume

$$S(t) = \sigma(\omega\, t) + \mathcal{E}(t) \tag{A 16}$$

where σ is a known function.

Then,

$$E\left[\hat{\omega} - \omega\right]^2 \geq \frac{\Psi\,\omega^3}{2}\left[\int_o^{2\pi N}\left[\theta\,\sigma'(\theta)\right]^2 d\,\theta\right]^{-1} \qquad \text{(A 17)}$$

$$= \frac{\Psi}{2}\,\frac{2\pi}{T}^3\left[\int_o^{2\pi}\left[\theta\,\sigma'(N\,\theta)\right]^2 d\,\theta\right]^{-1}$$

Now, finally, consider the case where $\sigma(\theta)$ is considered to be partly known. The formulation would be

$$S(t) = \sigma_a(\omega\,t) + \alpha\,\sigma_b\left[\omega(t - t_o)\right] + \varepsilon(t) \qquad \text{(A 18)}$$

where σ_a is completely unknown; σ_b is known; and α, ω, t_o are unknown.

It turns out, however, that if the equation for $E\left[\hat{\omega} - \omega\right]^2$ is applied in this case, without any further assumption of a-priori knowledge, the answer comes out exactly the same as for the case first treated, that is, the case where the total function σ is entirely unknown. This is also true if α and t_o are assumed known. The reason is, of course, that once σ_a is considered to be completely unknown, that is, that the coefficients in the expansion of σ_a relative to $\left\{P_i(\theta)\right\}$ can be anything, this is equivalent to saying that the expansion coefficients of $\sigma_a + \sigma_b$ can be anything.

Thus, it turns out that the problem where some portion of σ is known exactly cannot be properly formulated, in such a manner as to reflect the benefit of such knowledge, without associating significantly non-uniform a-priori statistics with the unknown part of σ.

For simplicity of exposition, assume that

$$S(t) = \sigma_a(\omega t) + \sigma_b(\omega t) + \mathcal{E}(t) \tag{A 19}$$

where

$$\sigma_a(\theta) = \sum_{i=1}^{\infty} \alpha_i P_i(\theta) \tag{A 20}$$

$$\sigma_b(\theta) = \sum_{i=1}^{\infty} \beta_i P_i(\theta)$$

and where all β_i are known a-priori.

Suppose that no a-priori statistics are associated with ω. However, suppose $\sigma_a(\theta)$ is known to be a random process having known mean and covariance function over $(0, 2\pi)$. Since the mean of $\sigma_a(\theta)$ is assumed known, we can replace $\sigma_a(\theta)$ by $\sigma_a(\theta) - \overline{\sigma_a(\theta)}$ and assume the mean to be zero.

Now, $\{P_i(\theta)\}$ has heretofore been considered a completely arbitrary complete orthonormal set in 0, 2π. At this point, we will make the following specific choice of $\{P_i(\theta)\}$: the orthonormal eigenfunctions of the covariance function $\Phi_a(\theta, \theta')$ of the random process $\{\sigma_a(\theta)\}$ over $(0, 2\pi)$.

In that case, the set $\{\alpha_i\}$ have a-priori statistics

$$\overline{\alpha_i} = o, \quad \text{all } i \tag{A 21}$$

$$\overline{\alpha_i \, \alpha_j} = A_i^2 \, \delta_{ij} \tag{A 22}$$

where A_i^2 are related to the eigenvalues of the kernel $\Phi_a(\theta, \theta')$.

We now have the actual observations given by

$$S(t) = f(t, \omega, \alpha_1, \alpha_2, \ldots) + \varepsilon(t) \qquad \text{(A 23)}$$

$$f(t, \omega, \alpha_1, \alpha_2, \ldots) = \sum_{i=1}^{\infty} (\alpha_i + \beta_i) P_i(\omega t) \qquad \text{(A 24)}$$

β_i are all known constants.

The virtual observations are given by

$$\alpha_i^{(e)} = \alpha_i + \delta \alpha_i^{(e)} \qquad \text{(A 25)}$$

where the observed values of $\alpha_i^{(e)}$ are $\overline{\alpha_i} = 0$ and the errors $\delta \alpha_i^{(e)}$ have covariance matrix given by Eq. (A 22).

If we now apply the generalized least squares smoothing technique to the set of observations consisting of both the actual and the virtual observations, we obtain the result that $E\left[\hat{\omega} - \omega\right]^2$ is given by Eqs. (A 6),, (A 9), (A 10), (A 12), and a modified version of Eq. (A 11):

$$B_{ij} = \frac{2}{\Psi}\left[\frac{N}{\omega} + A_i^{-2}\right] \delta_{ij} \qquad (i, j > 1) \qquad \text{(A 11a)}$$

Also, in applying these formulas,

$$\sigma(\theta) = \sigma_a(\theta) + \sigma_b(\theta) \qquad \text{(A 26)}$$

Caution must be used in interpreting this result, since it actually amounts to getting a lower bound on $E\left[\hat{\omega} - \omega\right]^2$ relative to the fictitious statistical ensemble of the actual and virtual observations. As pointed out in Section II, the statement that this can be considered equivalent to $E\left[\hat{\omega} - \omega\right]^2$ relative to the original statistical ensemble (defined by the statistics of $\{e(t)\}$ and the a-priori statistics of $\{\sigma(\theta)\}$ has been proved only if certain first order expansions are valid. The precise conditions under which the result stated is valid, for the case where non-uniform a-priori statistics are associated with a portion of σ, have not been investigated.

REFERENCES

1. Kelly, E. J., I. S. Reed, and W. Root, "The Detection of Radar Echoes in Noise: I and II", J. Soc. Indust. Appl. Math., Vol. 8, No. 2, June 1960, pp. 309-341; and Vol. 8, No. 3, September 1960, pp. 481-505.

2. Swerling, P., "Parameter Estimation Accuracy Formulas", IEEE Trans. Info. Theory, Vol. IT-10, No. 4, October 1964.

3. Swerling, P., "Maximum Angular Accuracy of a Pulsed Search Radar", Proc. IRE, Vol. 44, No. 9, September 1956, pp. 1146-1155.

4. Swerling, P., "Parameter Estimation for Waveforms in Additive Gaussian Noise", J. Soc. Indust. Appl. Math., Vol. 7, No. 2, June 1959, pp. 152-166.

5. Woodward, P. M., Probability and Information Theory with Applications to Radar, McGraw-Hill Book Co., New York, 1953.

6. Middleton, D. and D. Van Meter, "Detection and Extraction of Signals in Noise from the Point of View of Statistical Decision Theory: II", J. Soc. Indust. Appl. Math., Vol. 4, No. 2, June 1956, pp. 86-119.

7. Wiener, N., The Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications, John Wiley and Sons, New York.

8. Miller, K. S. and L. A. Zadeh, "Solution of an Integral Equation Occurring in the Theories of Prediction and Detection", IRE Trans. Info. Theory, Vol. IT-2, No. 2, June 1956, pp. 72-75.

9. Swerling, P., "First Order Error Propagation in a Stagewise Smoothing Procedure for Satellite Observations", Journal of the Astronautical Sciences, Vol. 6, No. 3, Autumn 1959, pp. 46-52.

10. Swerling, P., A Proposed Stagewise Differential Correction Procedure for Satellite Tracking and Prediction, The RAND Corporation, P-1292, January 8, 1958.

11. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems", Trans. ASME, Series D, Journal of Basic Engineering, Vol. 82, No. 1, March 1960, pp. 35-45.

12. Kalman, R. E. and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory", Trans. ASME, Series D, Journal of Basic Engineering, March 1961, pp. 95-108.

13. Blum, M., "A Stagewise Parameter Estimation Procedure for Correlated Data", Numerische Mathematik 3, 1961, pp. 202-208.

14. Swerling, P., "Optimum Linear Estimation for Random Processes as the Limit of Estimates Based on Sampled Data", IRE Wescon Convention Record, Part 4, 1958, pp. 158-163.

15. Bryson, A. F. and D. E. Johansen, "Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise," IEEE Trans. on Automatic Control, Vol. AC-10, No. 1, January 1965, pp. 4-10.

16. Swerling, P., "Classes of Signal Processing Procedures Suggested by Exact Minimum Mean Square Error Procedures," Journ. Soc. Indust. Appl. Math., (to appear).